The Road to Billion Transistor Processor Chips in the Near Future

Dileep Bhandarkar
Roger Golliver

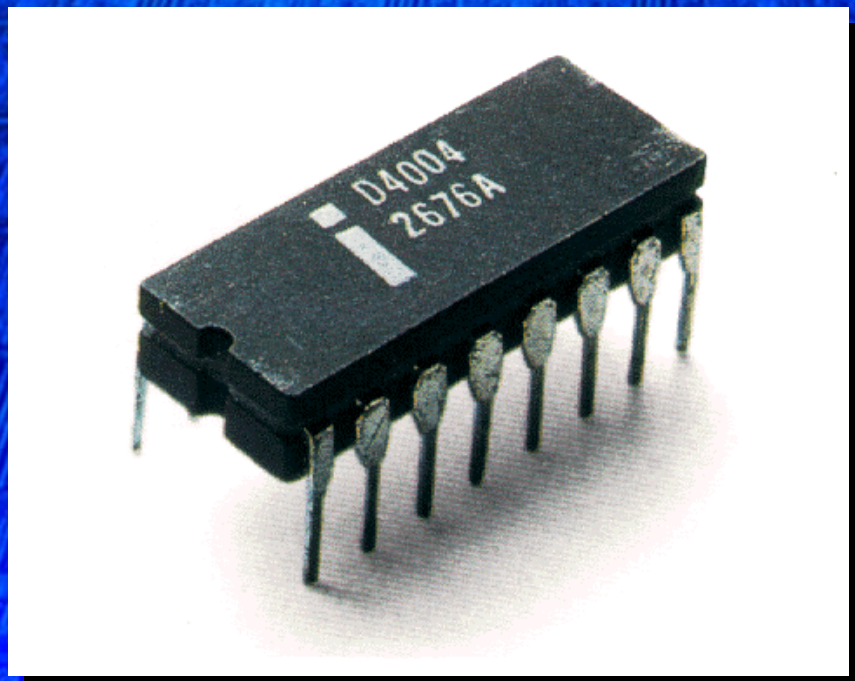
Intel Corporation

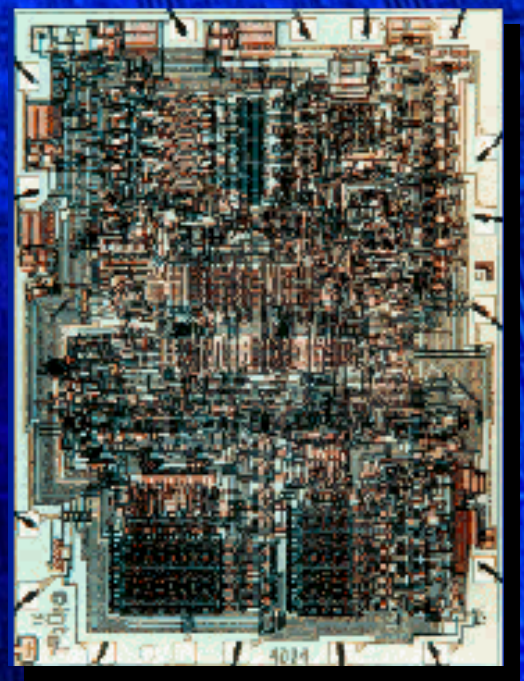
April 22th, 2003

Outline

- Semiconductor Technology Evolution
- Moore's Law Video
- Parallelism in Microprocessors Today
- Multiprocessor Systems
- The Path to Billion Transistors
- Summary

Birth of the Revolution -- The Intel 4004





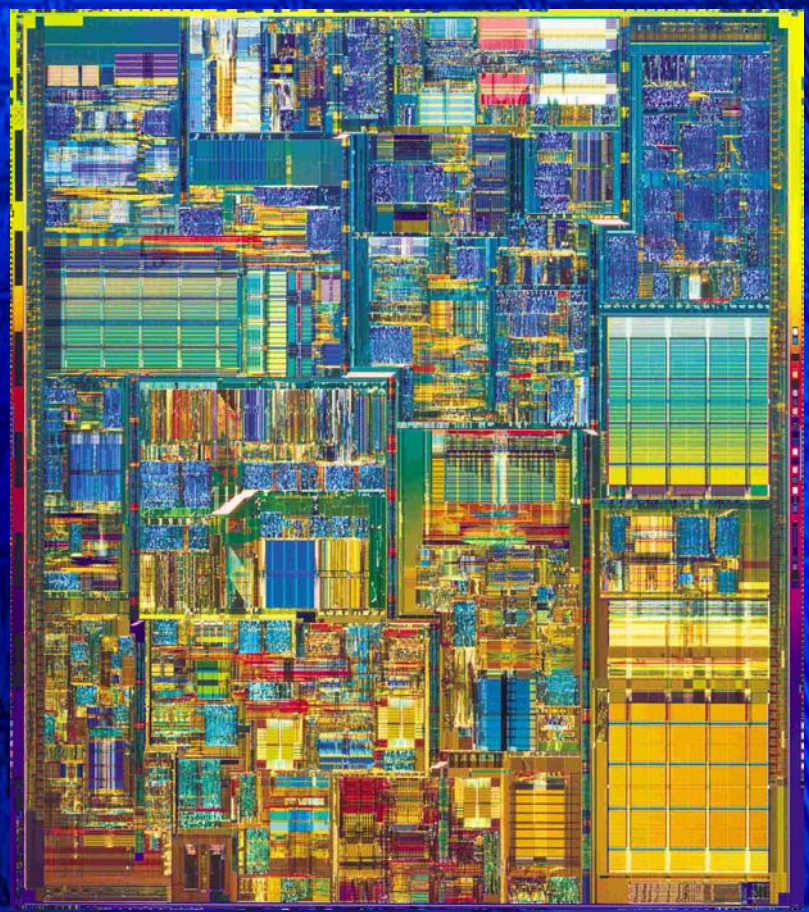
Introduced November 15, 1971
108 KHz, 50 KIPs , 2300 10μ transistors

intel

2001 – Pentium® 4 Processor

Introduced November 20, 2000

@1.5 GHz core, 400 MT/s bus
42 Million 0.18µ transistors



August 27, 2001

@2 GHz, 400 MT/s bus
640 SPECint_base2000*
704 SPECfp_base2000*

intel

Source: http://www.specbench.org/cpu2000/results/

30 Years of Progress

- 4004 to Pentium® 4 processor
  - Transistor count: 20,000x increase
  - Frequency: 20,000x increase
  - 39% Compound Annual Growth rate

intel

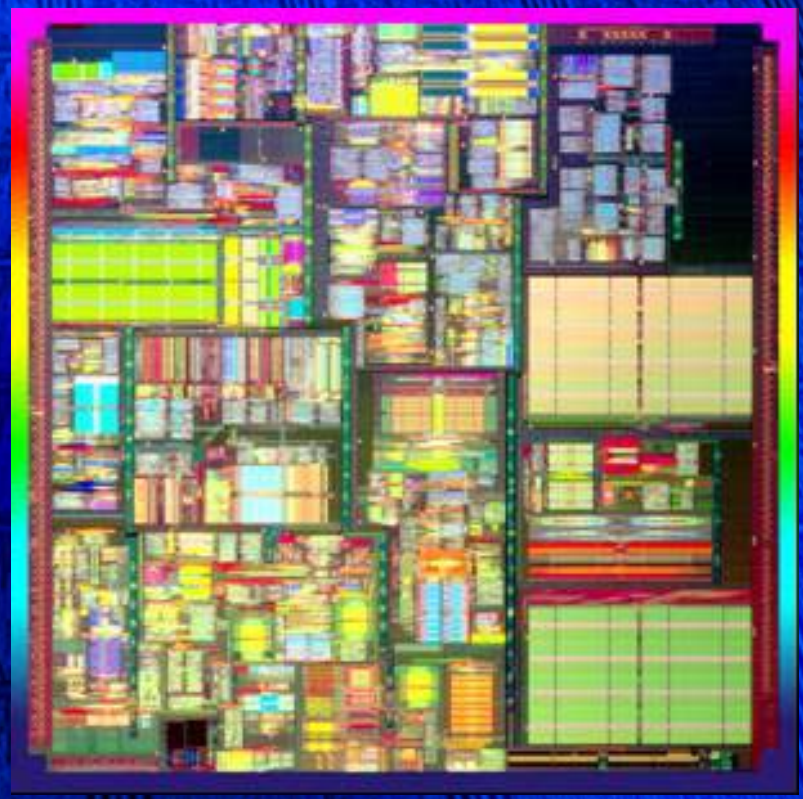
2002 – Pentium® 4 Processor

November 14, 2002

@3.06 GHz, 533 MT/s bus

1099 SPECint_base2000*
1077 SPECfp_base2000*

55 Million 130 nm process

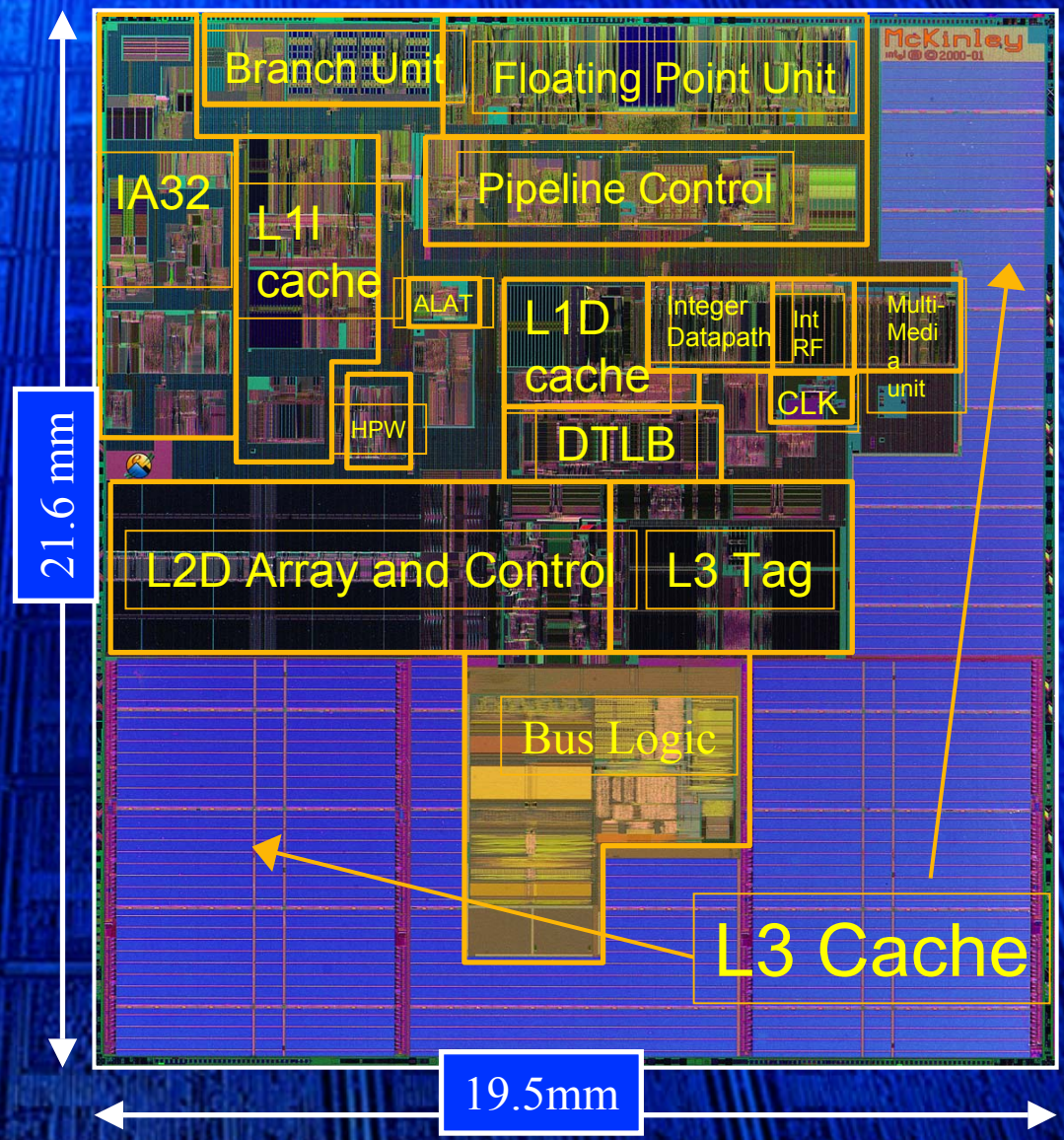


intel

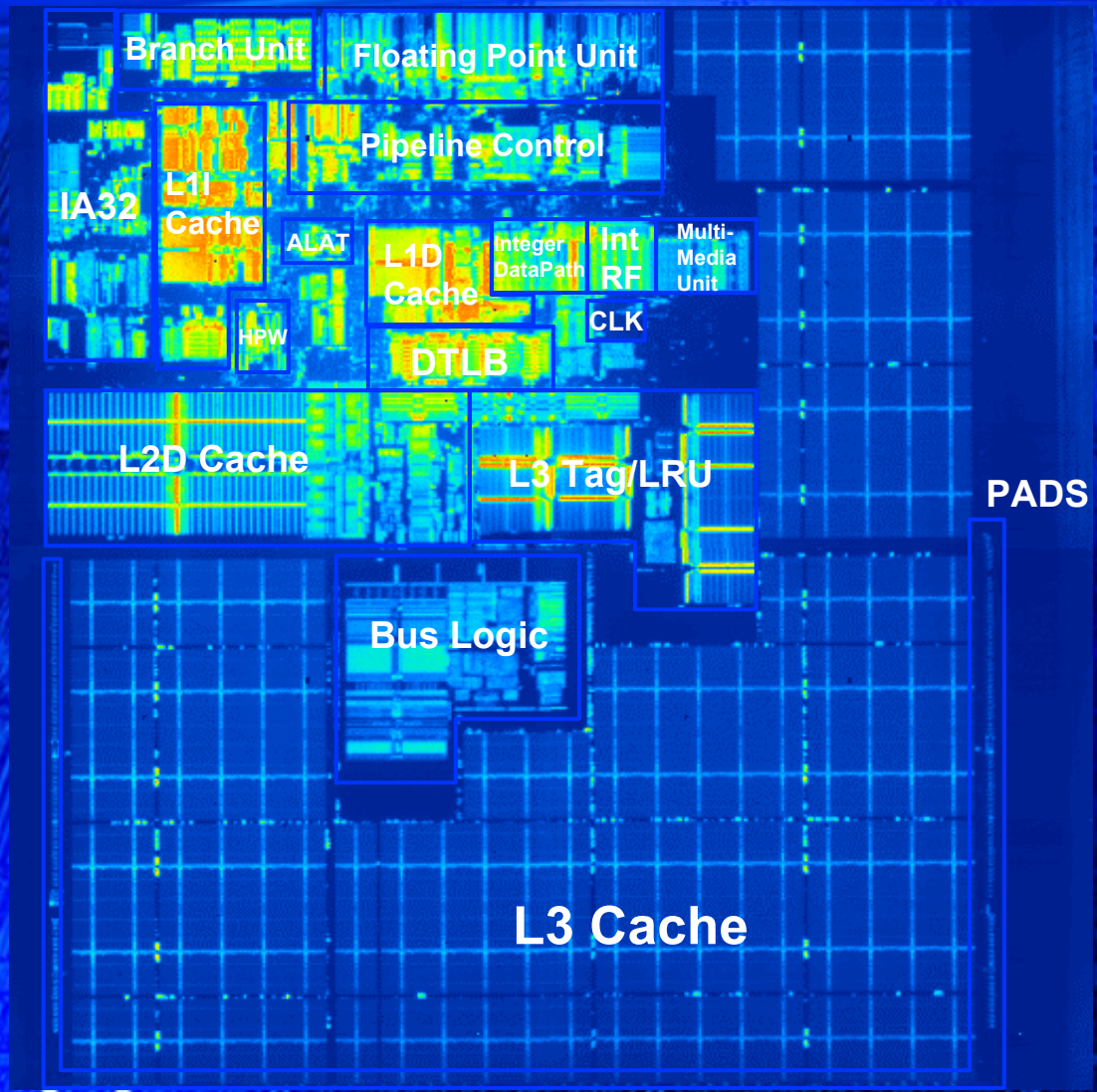
Source: http://www.specbench.org/cpu2000/results/

Itanium® 2 Processor Overview

- .18µm bulk, 6 layer Al process
- 8 stage, fully stalled in-order pipeline
- Symmetric six integer-issue design
- IA32 execution engine integrated
- 3 levels of cache on-die totaling 3.3MB
- 221 Million transistors
- 130W @1GHz, 1.5V
- 421 mm^2 die
- 142 mm^2 CPU core



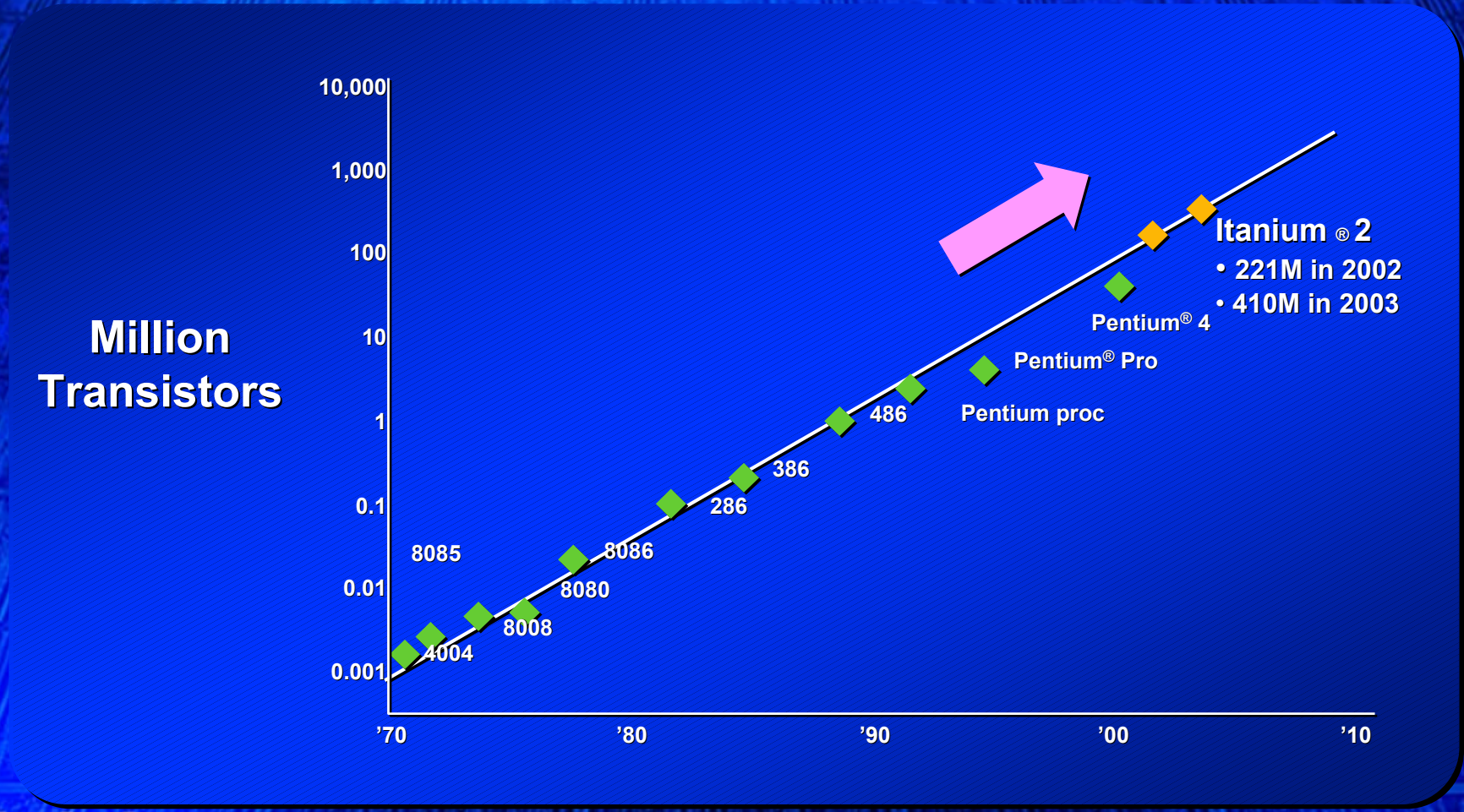
Madison Processor



- 130 nm process
- ~1.5 GHz
- 6 MB L3 Cache
- 24 way set associative
- ~130W
- 410 M transistors
- 374 square mm

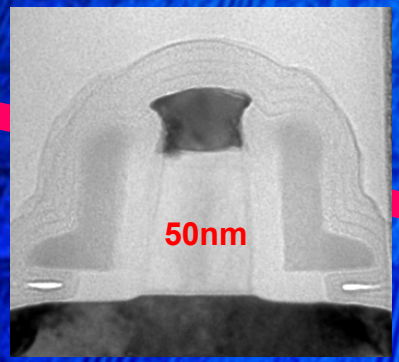
intel

Continuing at this Rate by End of the Decade

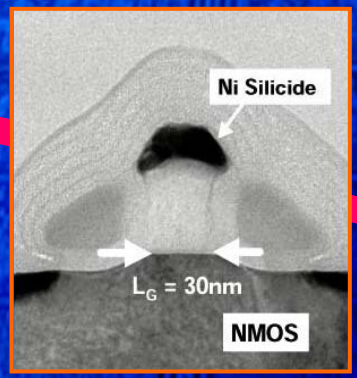


Billion Transistors possible within 4 years

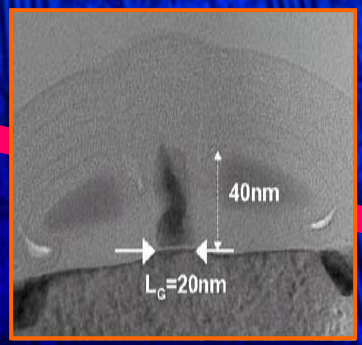
Nanotechnology Advancements



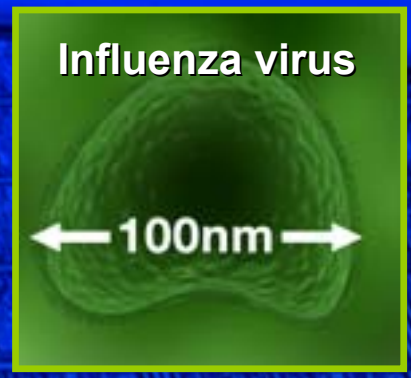
90nm Node
Lgate = 50nm
Production - 2003

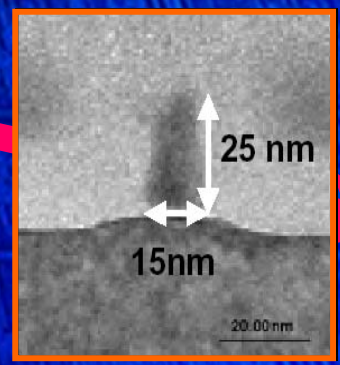


65nm Node
Lgate = 30nm
Production - 2005



45nm Node
Lgate = 20nm
Production - 2007





30nm Node
Lgate = 15nm
Production - 2009

intel

Process Advancements Fulfill Moore's Law

Source: Intel

Semiconductor Manufacturing Process Evolution

	Actual					Forecast	
Process name	P852	P854	P856	P858	Px60	P1262	P1264
Production	1993	1995	1997	1999	2001	2003	2005
Generation	0.50	0.35	0.25	0.18µm	130 nm	90 nm	65 nm
Gate Length	0.50	0.35	0.20	0.13	<70 nm	<50 nm	<35 nm
Wafer Size (mm)	200	200	200	200	200/300	300	300

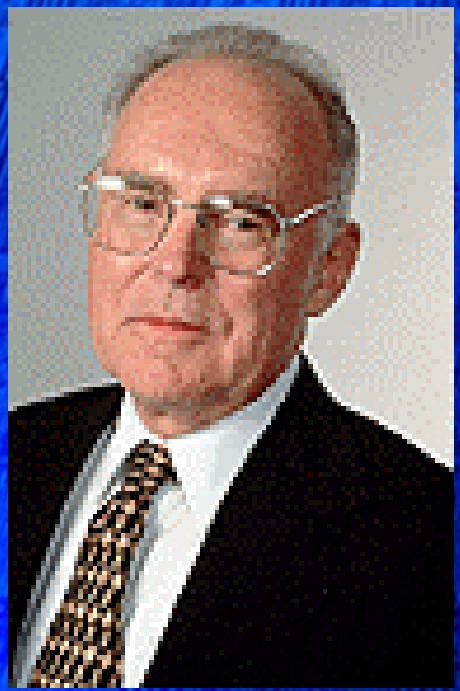
New generation every 2 years

intel

Outline

- Semiconductor Technology Evolution
- Moore's Law Video
- Parallelism in Microprocessors Today
- Multiprocessor Systems
- The Path to Billion Transistors
- Summary

Moore's Law



The experts look ahead

Cramming more components onto integrated circuits

With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore

Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

The future of integrated electronics is the future of electronics itself. The advantages of integration will bring about a proliferation of electronics, pushing this science into many new arenas.

Integrated circuits will lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment. The electronic wristwatch needs only a display to be feasible today.

But the biggest potential lies in the production of large systems. In telephone communications, integrated circuits in digital filters will separate channels on multiplex equipment. Integrated circuits will also switch telephone circuits and perform data processing.

Computers will be more powerful, and will be organized in completely different ways. For example, memories built of integrated electronics may be distributed throughout the machine instead of being concentrated in a central unit. In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units. Machines similar to those in existence today will be built at lower costs and with faster turn-around.

Present and future

By integrated electronics, I mean all the various technologies which are referred to as microelectronics today as well as any additional ones that result in electronics functions supplied to the user as irreducible units. These technologies were first investigated in the late 1950's. The object was to miniaturize electronics equipment to include increasingly complex electronic functions in limited space with minimum weight. Several approaches evolved, including microassembly techniques for individual components, thin-film structures and semiconductor integrated circuits.

Each approach evolved rapidly and converged so that each borrowed techniques from another. Many researchers believe the way of the future to be a combination of the various approaches.

The advocates of semiconductor integrated circuitry are already using the improved characteristics of thin-film resistors by applying such films directly to the active semiconductor substrate. Those advocating a technology based upon films are developing sophisticated techniques for the attachment of active semiconductor devices to the passive film arrays.

Both approaches have worked well and are being used in equipment today.

The author

Dr. Gordon E. Moore is one of the new breed of electronic engineers, schooled in the physical sciences rather than in electronics. He earned a B.S. degree in chemistry from the University of California and a Ph.D. degree in physical chemistry from the California Institute of Technology. He was one of the founders of Fairchild Semiconductor and has been director of the research and development laboratories since 1959.

Electronics, Volume 38, Number 8, April 19, 1965

Outline

- Semiconductor Technology Evolution
- Moore's Law Video
- Parallelism in Microprocessors Today
- Multiprocessor Systems
- The Path to Billion Transistors
- Summary

Different Forms of Parallelism

- Within a processor core
  - multiple issue processors with lots of execution units
  - wider superscalar
  - explicit parallelism
- Multiple processors on a chip
  - Hardware Multi Threading
  - Multiple cores

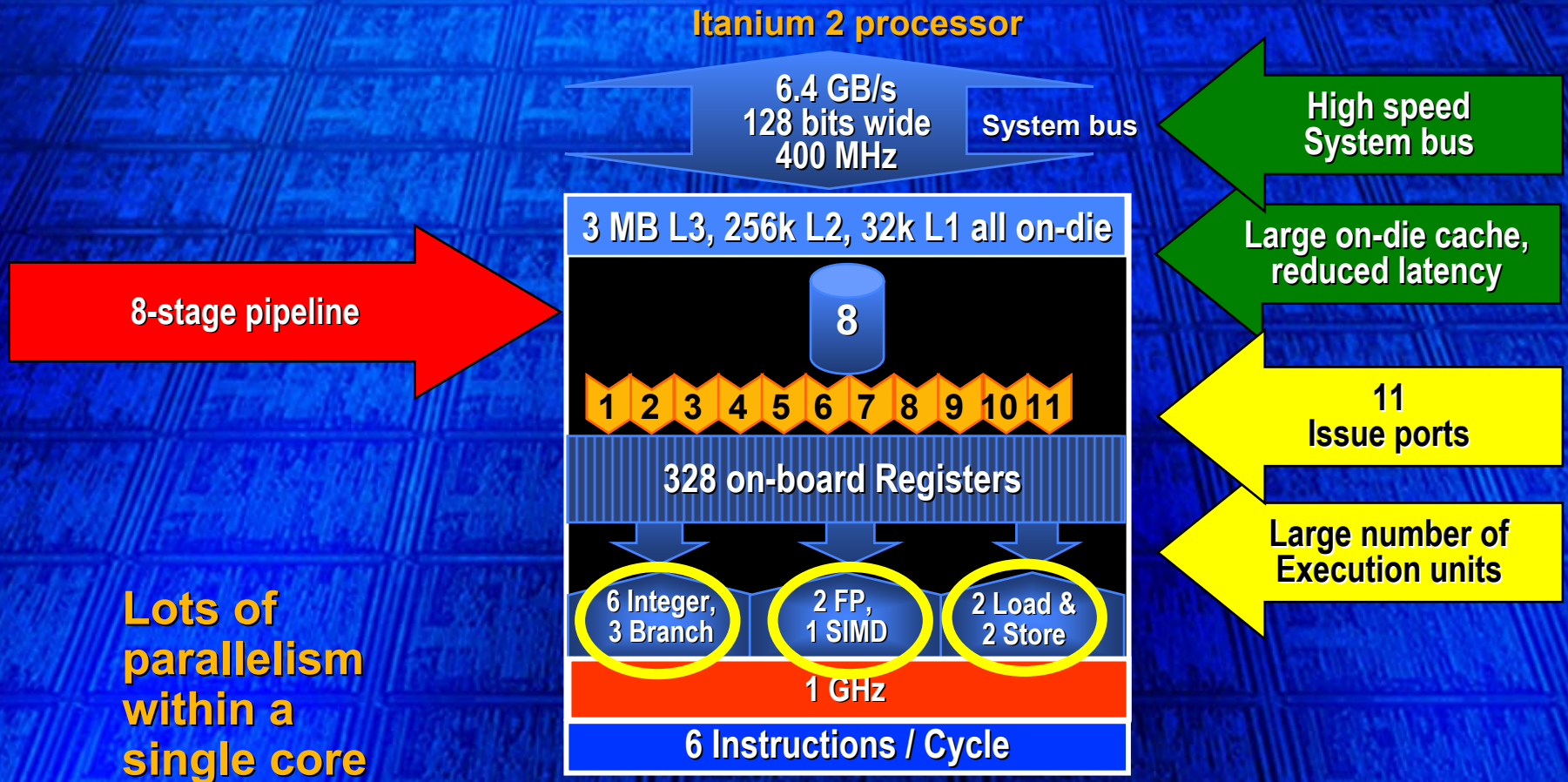
EPIC Architecture Features

Explicitly Parallel Instruction Computing

- Enable wide execution by providing processor implementations that compiler can take advantage of
- Performance through parallelism
  - Multiple execution units and issue ports in parallel
  - 2 bundles (up to 6 Instructions) dispatched every cycle
- Massive on-chip resources
  - 128 general registers, 128 floating point registers
- 64 predicate registers, 8 branch registers
  - Exploit parallelism
  - Efficient management engines (register stack engine)
- Provide features that enable compiler to reschedule programs using advanced features (predication, speculation, software pipelining)
- Enable, enhance, express, and exploit parallelism

intel

Itanium® 2 Processor Architecture



Lots of parallelism within a single core

Itanium 2 processor
221 million transistors total
25 million in CPU core logic

intel

High Performance Design Space

With Each Process Generation

- Frequency increases by about 1.5X
- Vcc will scale by only ~0.8
- Active power will scale by ~0.9
- Active power density will increase by ~30-80%
- Leakage power will make it even worse

Doubling performance requires more than 4 times the transistors

Challenges: Power and Design Complexity

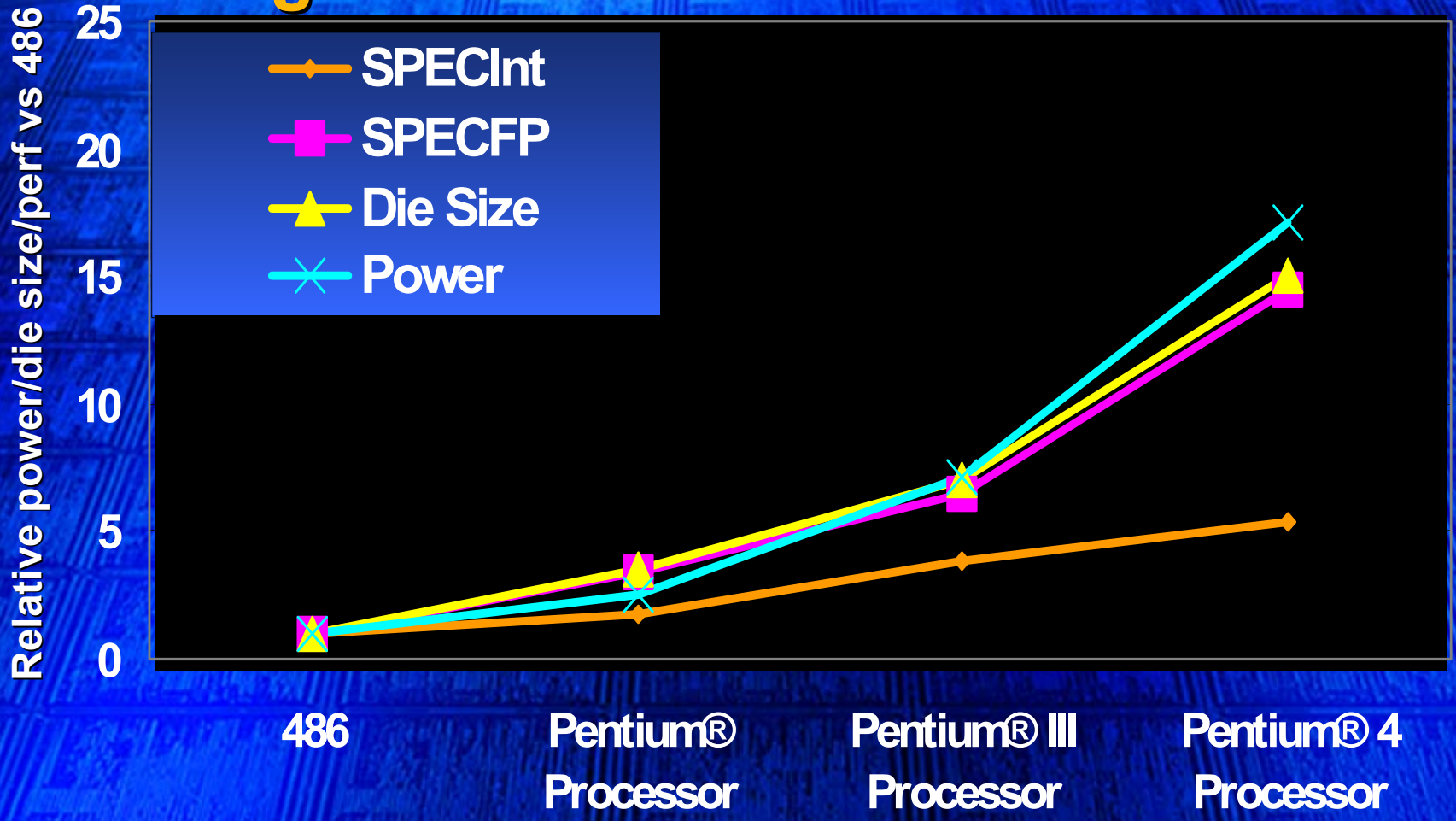
Power & Performance Tradeoffs

- Throughput Perf α SQRT(Frequency)
- But Power α Capacitance * Voltage^2 * Frequency
  - Frequency α Voltage
- Power increases non-linearly with Frequency
- Throughput Performance can be achieved at Lower Power with multiple low power cores
- Tradeoff between single stream and throughput performance for a constant power envelop

intel

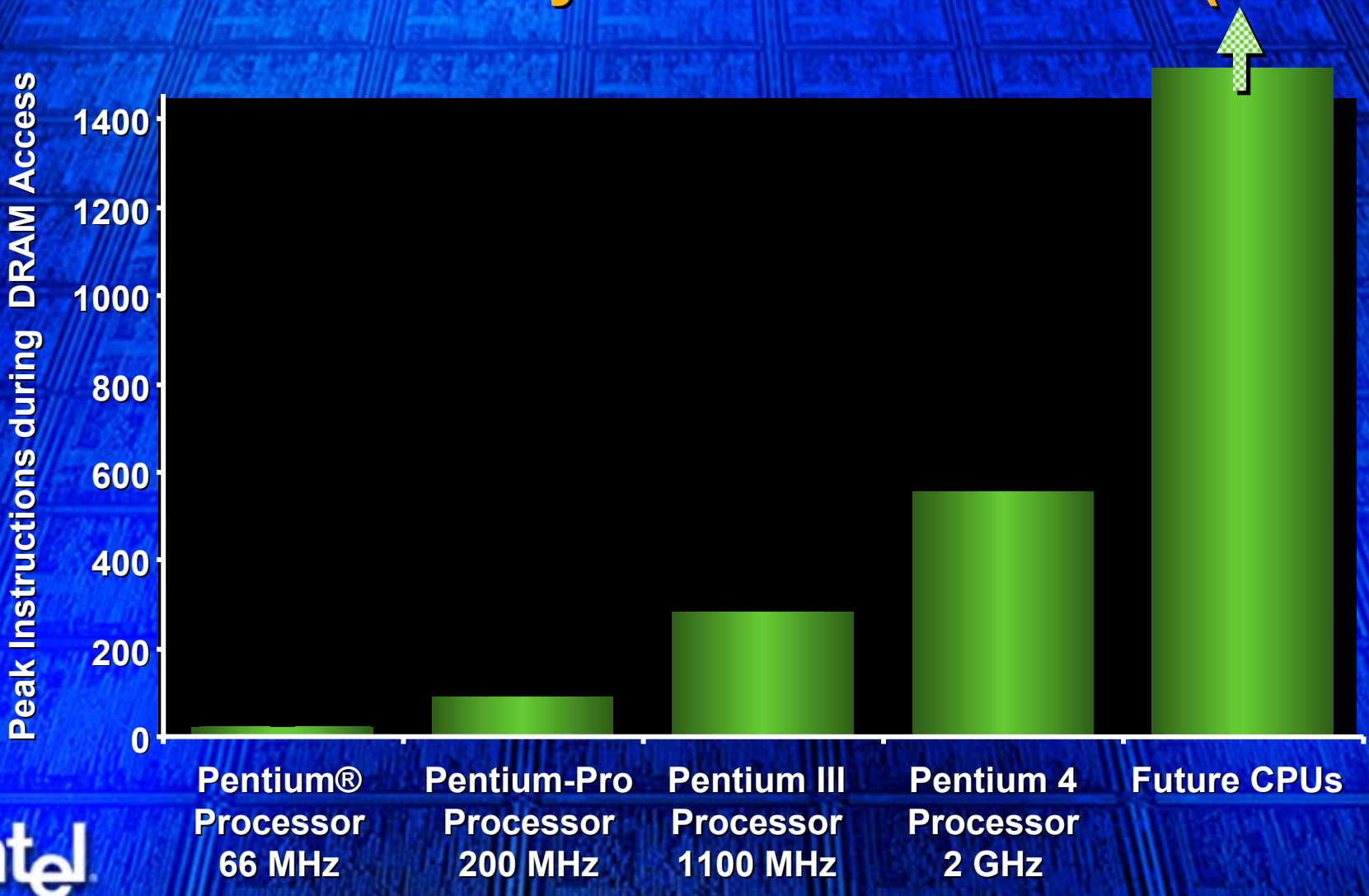
Can optimize for either Throughput or Single Stream

Single-stream Performance vs Costs

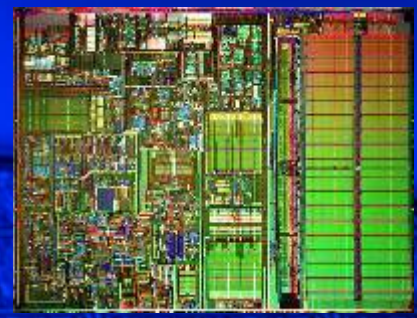


Die Size and Power have increased with Scalar Performance

Long Latency DRAM Accesses: Needs Memory Level Parallelism (MLP)



Multithreading



- Introduced on Intel® Xeon™ Processor MP
- Two logical processors for < 5% additional die area
- Executes two tasks simultaneously
  - Two different applications
  - Two threads of same application
  - Overlap execution behind memory access
- CPU maintains architecture state for two processors
  - Two logical processors per physical processor
- Power efficient performance gain
- 20-30% performance improvement on many throughput oriented workloads

intel

HyperThreading Technology: What was added to Intel Xeon MP Processor?

Instruction Streaming Buffers

Next Instruction Pointer

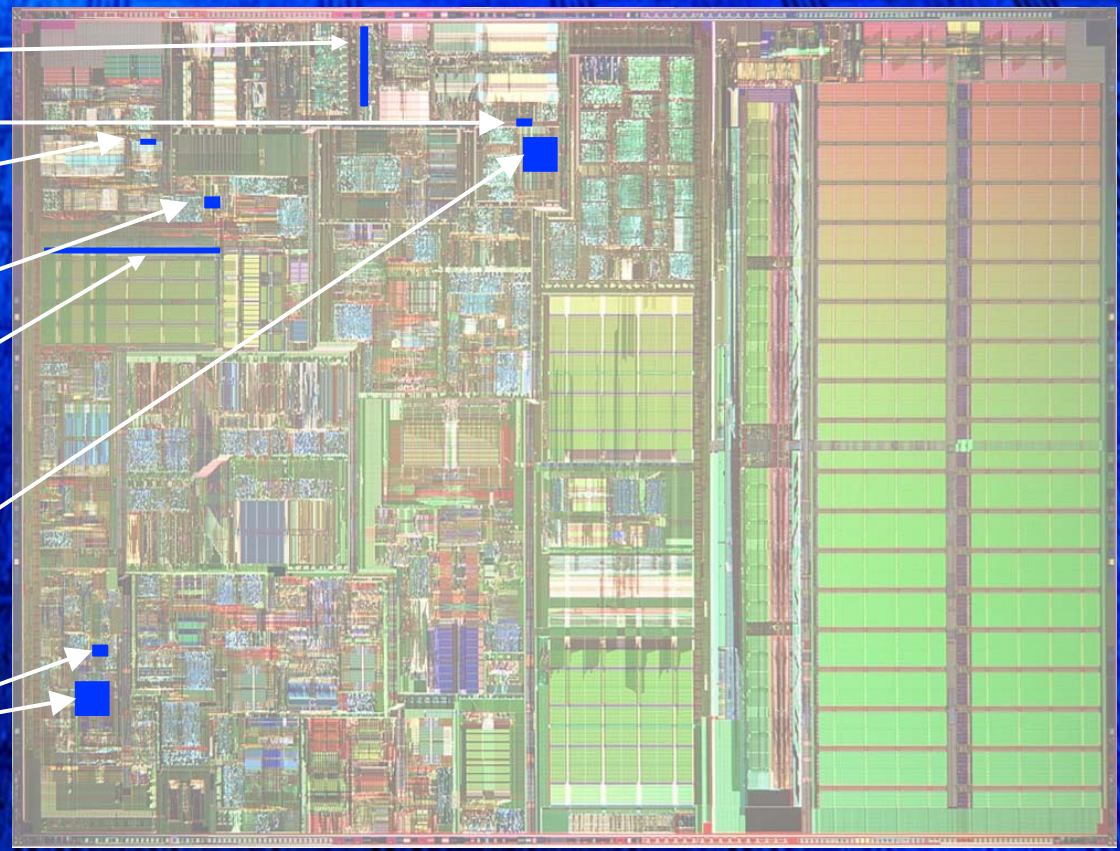
Return Stack Predictor

Trace Cache Next IP

Trace Cache Fill Buffers

Instruction TLB

Register Alias Tables

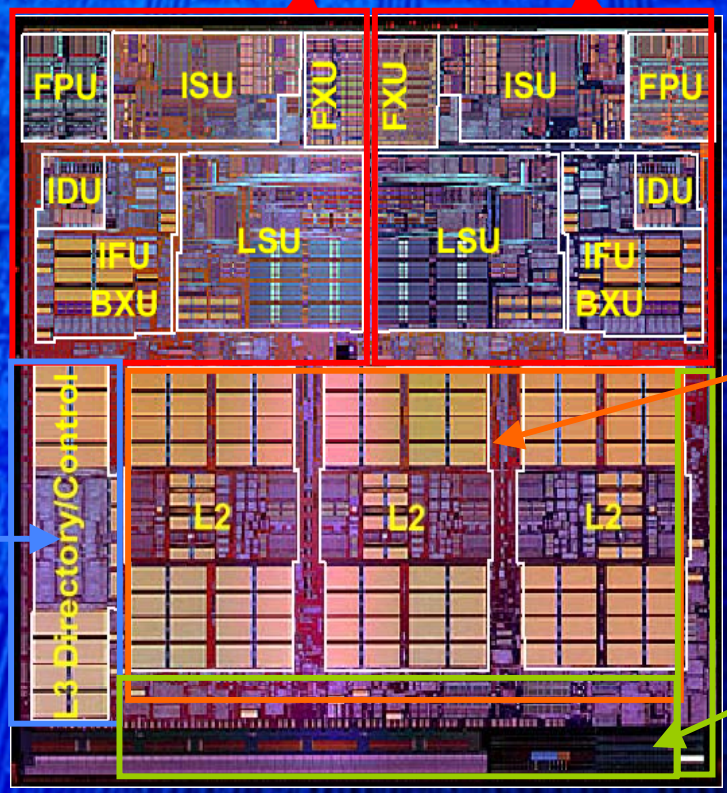


<5% die size (& max power), up to 30% performance increase

intel

IBM Power4 Dual Processor on a Chip

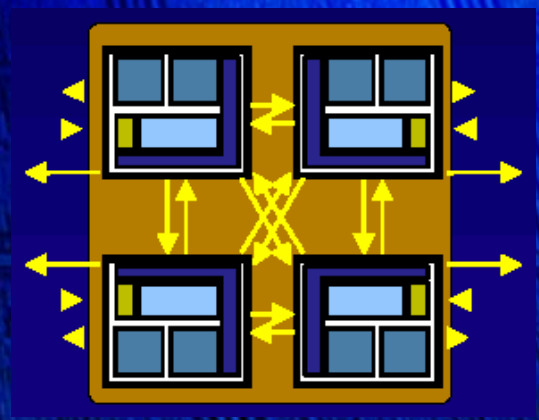
Two cores (~30M transistors each)



L3 & Mem Controller: L3 tags on-die for full-speed coherency checks

Large Shared L2: Multi-ported: 3 independent slices

Chip-to-Chip & MCM-to-MCM Fabric: Glueless SMP

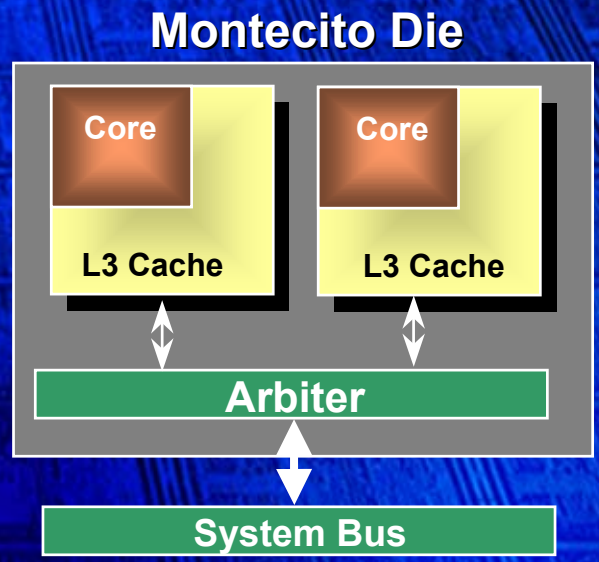


intel

*Other names and brands may be claimed as the property of others

Montecito Details

- Dual-core processor
- Each core with its own L3 cache
- Larger cache than previous processor
- Arbiter manages two cores as 1 bus interface
  - Enables same socket and protocol



intel

Outline

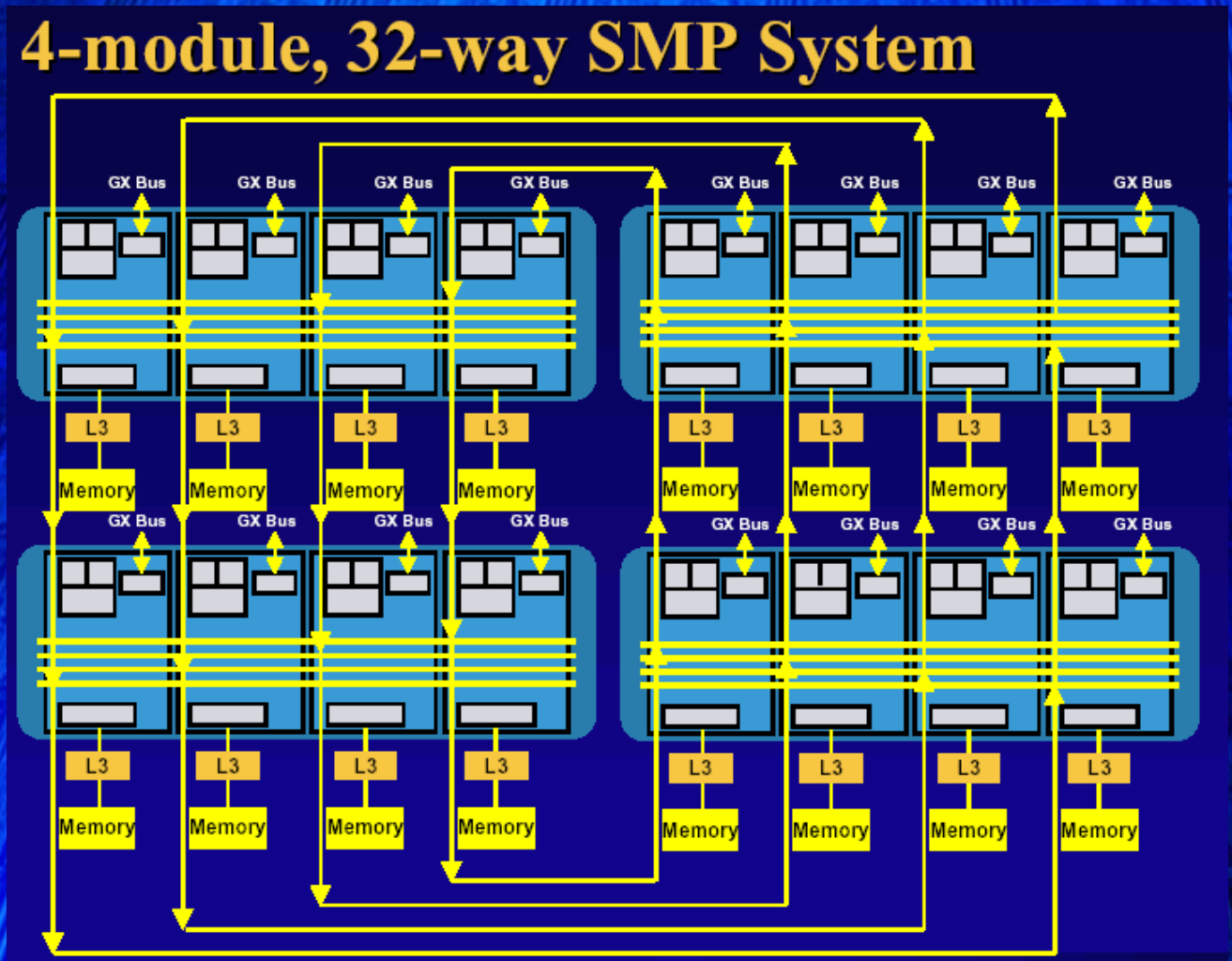
- Semiconductor Technology Evolution
- Moore's Law Video
- Parallelism in Microprocessors Today
- Multiprocessor Systems
- The Path to Billion Transistors
- Summary

Large Multiprocessor Systems

- 32 and 64 processor systems available today
- >400K transactions per minute
- >100 Linpack Gigaflops

Historically, system level features have migrated to microprocessor chips as transistor densities have increased

IBM eServer pSeries 690

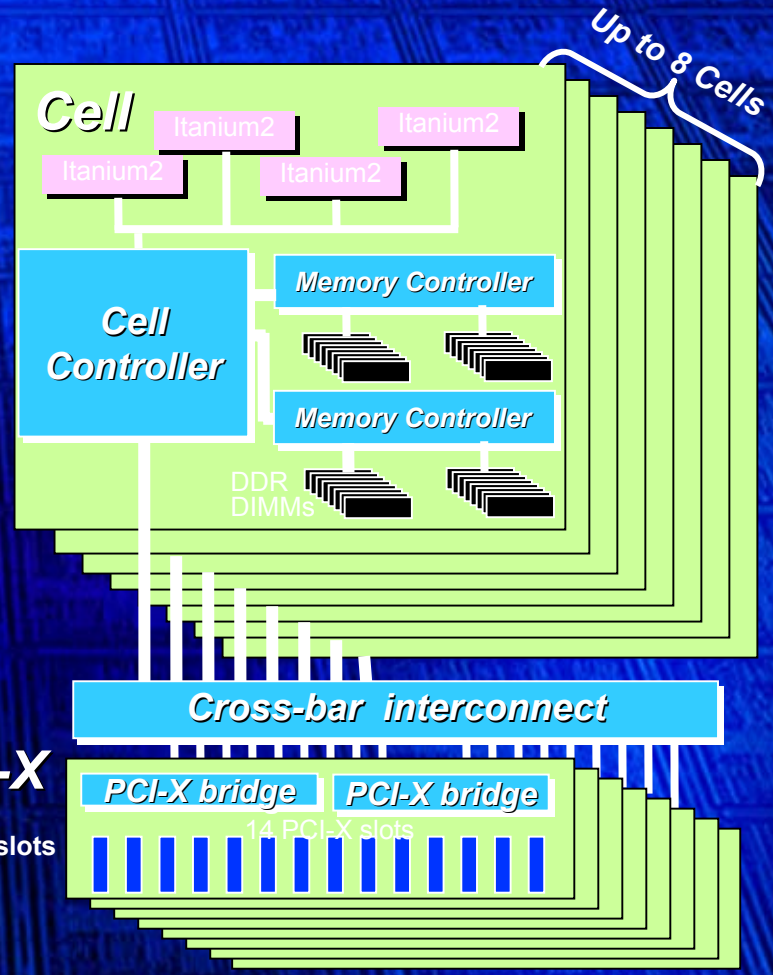


- 1.3 GHz Power4
- 8 to 32 CPU
- 8-way MCM @ $275,000**
- 403,255 tpmc @ $17.80 per tpmC***
- 95 Linpack Gflops

***http://www-132.ibm.com/content/home/store_IBMPublicUSA/en_US/eServer/pSeries/high_end/pSeries_highend.html
***Source: http://www.tpc.org/results/individual_results/IBM/IBMp690es_08142002.pdf
*Other names and brands may be claimed as the property of others

NEC Express5800/1320Xc SMP Server

- Up to 32 Itanium® 2 processors
- Up to 512GB memory (with 2GB DIMMs)
- Up to 112 PCI-X I/O slots
- Low latency and high bandwidth cross-bar interconnect
- Inter-cell memory interleaving
- ECC protected data transfer
- 433,107 tpmC @ $12.98 per tpmC**
- 101 Linpack GigaFlops
- 32 Processors + 512GB @ $2,126,090**



**http://www.tpc.org/results/individual_results/NEC/nec.express5800.1320xc.c5.030220.es.pdf

*Other names and brands may be claimed as the property of others

HPC Clusters

15 node dual 2.4GHz Pentium® Xeon® cluster

- Commercial Off The Shelf (COTS) components
  - Processors
  - Packaging
  - Interconnects
  - Operating systems

2,304 Intel® Xeon™ 2.4 GHz processors power this 5.69 TFlops supercomputer at Lawrence Livermore National Labs. It rates as the fifth fastest in the world.

intel

Outline

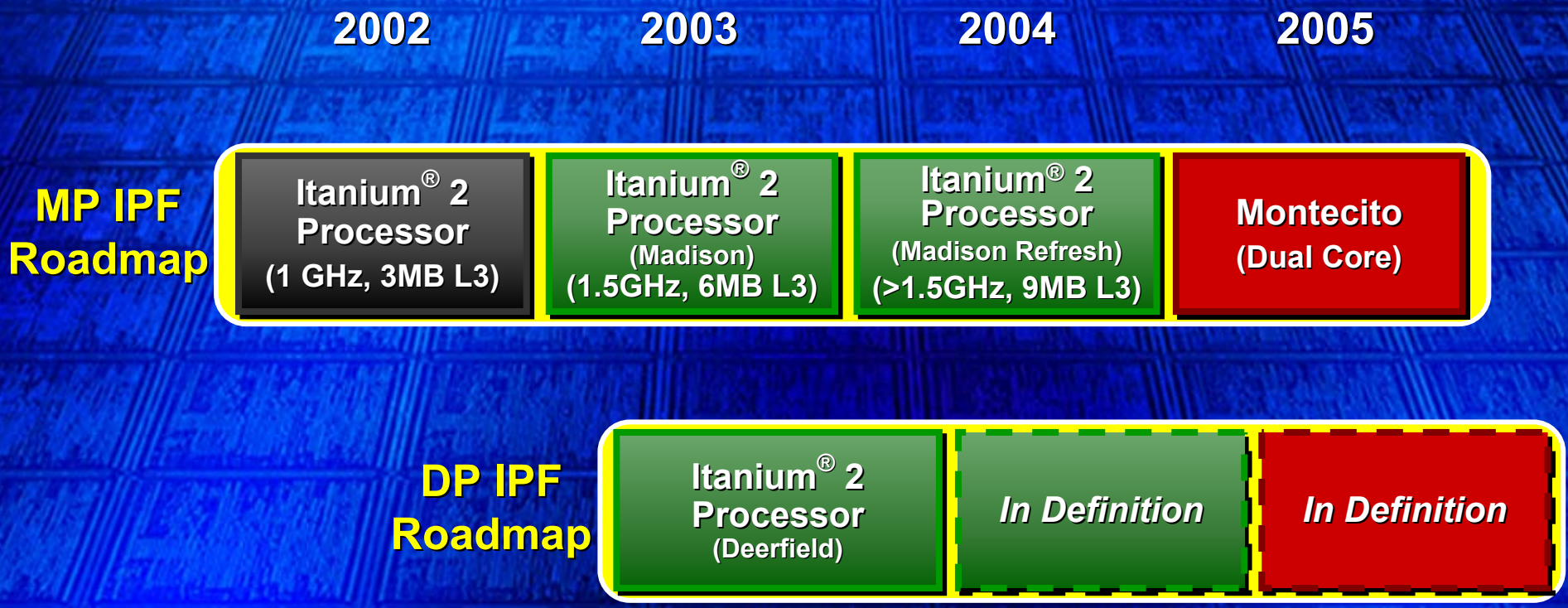
- Semiconductor Technology Evolution
- Moore's Law Video
- Parallelism in Microprocessors Today
- Multiprocessor Systems
- The Path to Billion Transistors
- Summary

Source: Intel estimates. All products, dates and information are preliminary and subject to change without notice.

Itanium® Processor Family Roadmap

Focus on Compatibility Preserves OEM and End-user Investments

Process
- 0.18 µm
- 0.13 µm
- 90 nm



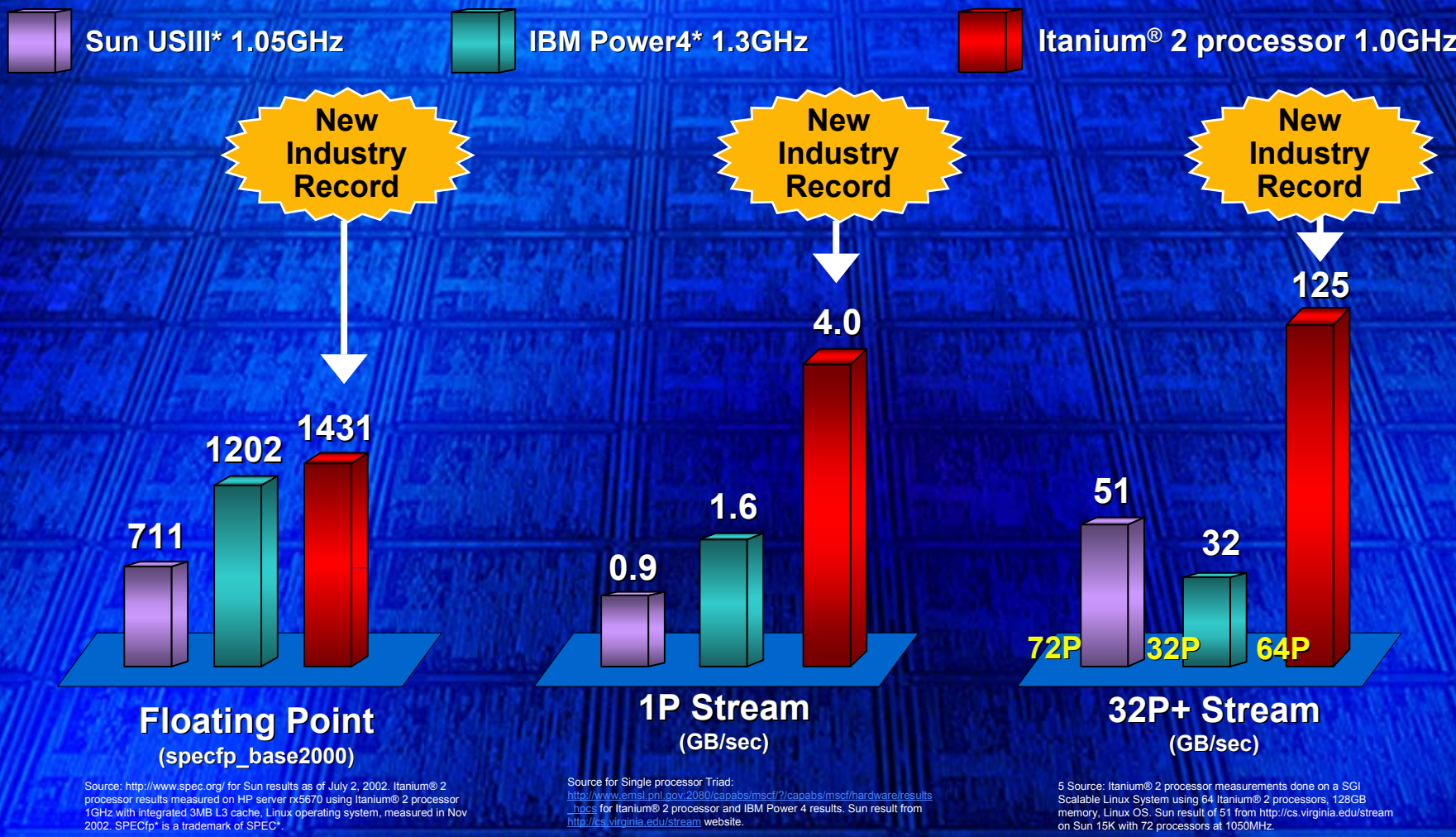
All dates specified are target dates, are provided for planning purposes only and are subject to change.

intel

Steady Performance Growth Over Time

2002 Vintage Itanium® 2 Processor

Record Setting Performance for Technical Computing



Source: http://www.spec.org/ for Sun results as of July 2, 2002. Itanium® 2 processor results measured on HP server rx5670 using Itanium® 2 processor 1GHz with integrated 3MB L3 cache, Linux operating system, measured in Nov 2002. SPECfp* is a trademark of SPEC*.

Source for Single processor Triad: http://www.emsl.pnl.gov:2080/capabs/mscf/?/capabs/mscf/hardware/results_hpcs for Itanium® 2 processor and IBM Power 4 results. Sun result from http://cs.virginia.edu/stream website.

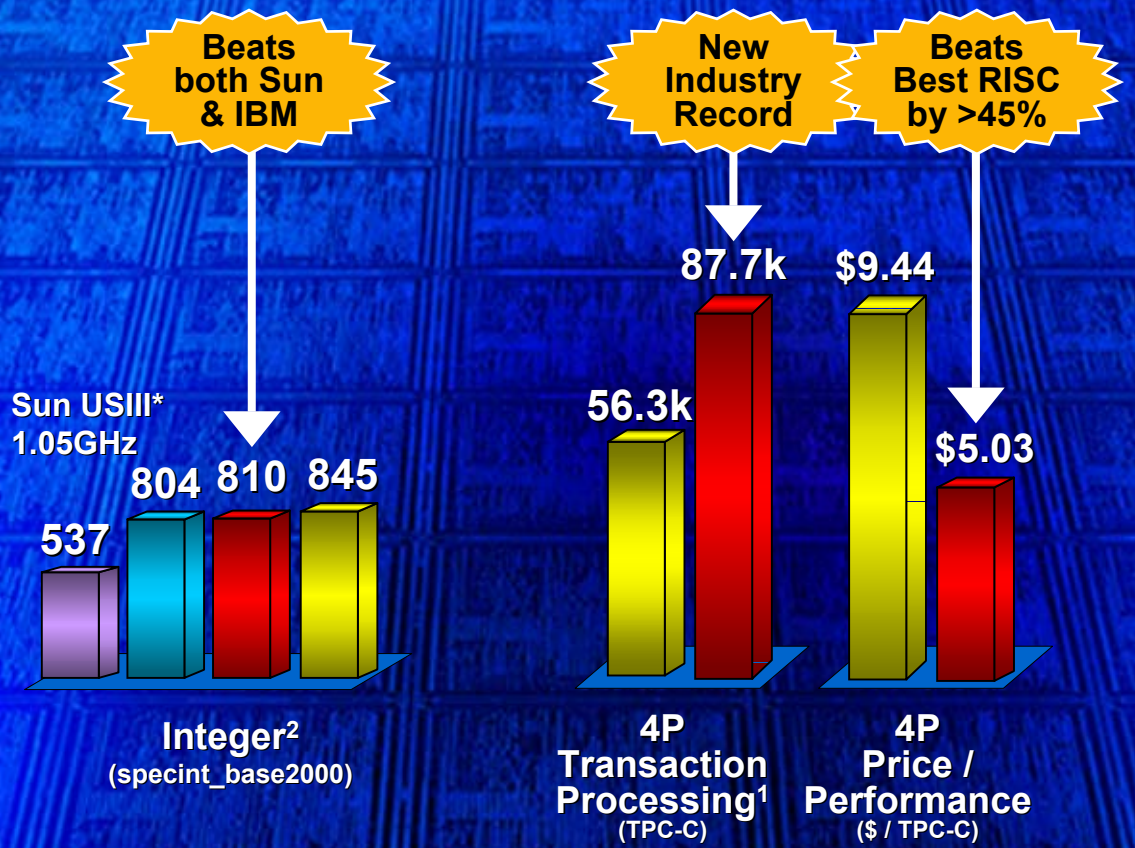
5 Source: Itanium® 2 processor measurements done on a SGI Scalable Linux System using 64 Itanium® 2 processors, 128GB memory, Linux OS. Sun result of 51 from http://cs.virginia.edu/stream on Sun 15K with 72 processors at 1050MHz.

Source: www,spec.org as of December 2002

Itanium® 2 Processor

Record Setting Performance for the Enterprise

Alpha* 1.25GHz IBM Power4* 1.3GHz Itanium® 2 Processor 1.0GHz



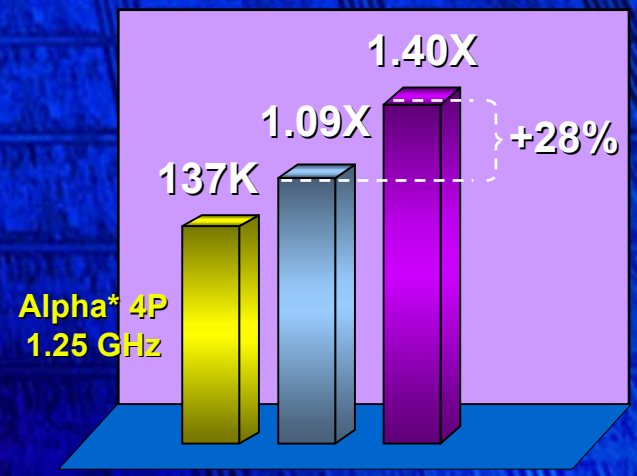
Data Sources: 1) TPC-C: www.tpc.org , 2) SPEC benchmarks: www.spec.org, 3) SAP SD www.sap.com//benchmark

2003 Itanium® 2 Processors

Madison / Deerfield Overview

- Extends Itanium® 2 processor architecture to .13u process technology
  - Production targets May - July 2003
- Delivers full platform compatibility
  - Socket and system bus compatible
  - Binary compatible with Itanium software
- Madison processor features
  - 1.5GHz, 6 MB L3 Cache
  - +30%-50% performance than Itanium® 2 processor

Pre-Production Madison Processor Performance Results – Sept. '02



4P SAP APO

Itanium® 2 processor, 1.0 GHz

Madison processor 1.3 GHz
(pre-production – Sept. '02)

Emphasis on compatibility preserves OEM and end-user investment in today's Itanium 2-based systems

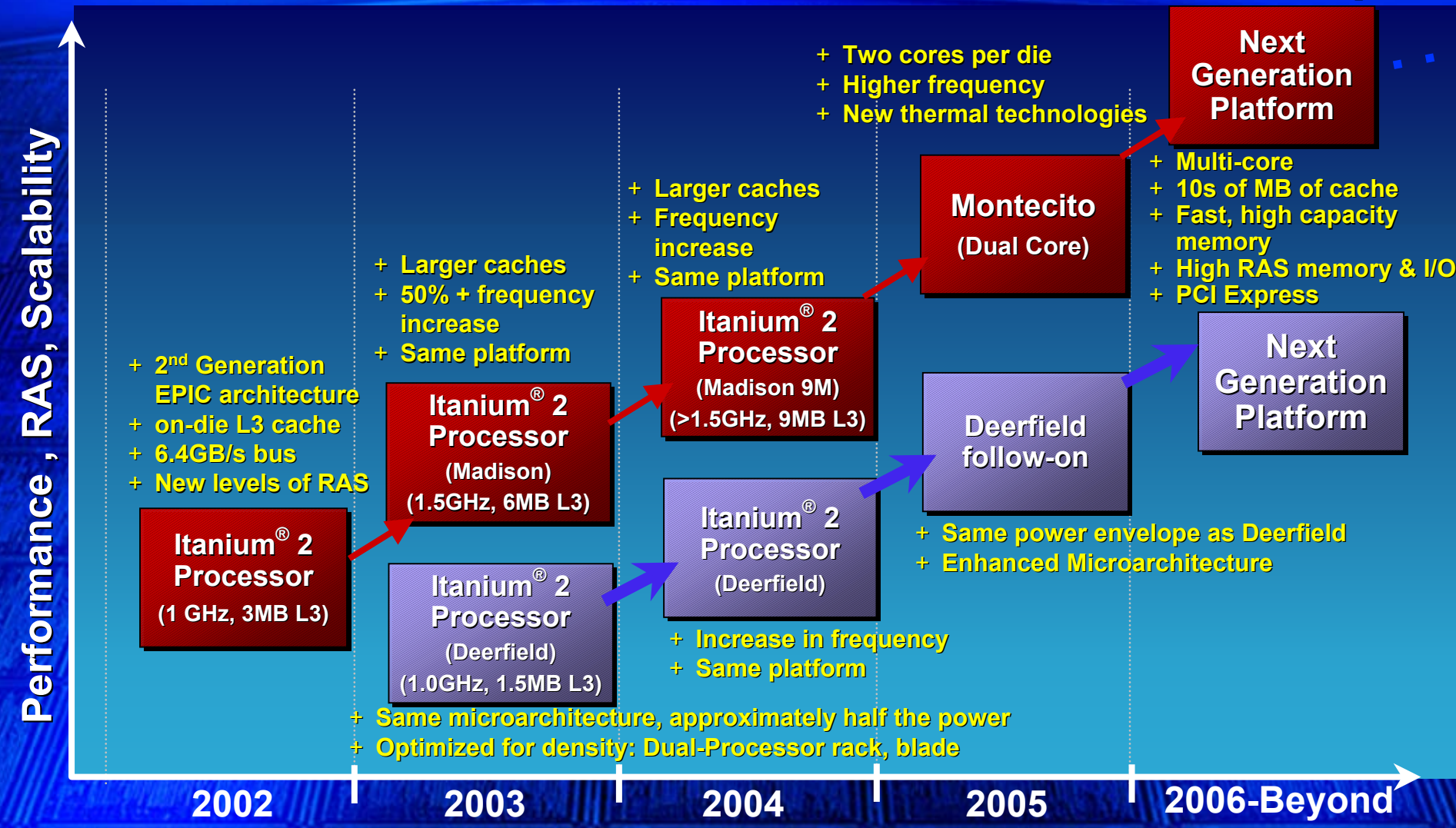
2 Source: Alpha results using HP Alphaserver ES45, 4P SMP, Alpha 2164C 1.25GHz, 16MB L2 cache, 32GB, APO rel. 3.0, Live Cache rel. 7.2.5.7, Oracle 9i Itanium® 2 processor and Madison numbers are preliminary lab results using HP server RX5670, 4 Itanium 2 processors 1.0GHz w/ 3M integrated L3 cache or 4 Madison 6M integrated L3 cache, 32 GB, APO rel. 3.0, Live Cache 7.4.2, Windows .Net Enterprise Server LE1.2 , SQL Server Enterprise Edition (64-bit) beta.

Itanium® 2 Processor (Deerfield) Summary

- Key features
  - 1.0GHz core frequency, 1.5MB L3 cache, DP only
  - 62W maximum power
- Compatibility
  - Based on the Itanium® 2 microarchitecture
  - Binary compatible with Itanium® processor software
  - Socket and system bus compatible with Itanium® 2-based DP systems
  - Intel® 870 chipset support
- Target market
  - Performance rack or blade servers and high-end workstations
- Schedule
  - Platform release target is 2H03, after Madison

Performance of today's Itanium® 2 processor 1.0 GHz at less than half the power

Intel® Itanium® Architecture Processor Roadmap



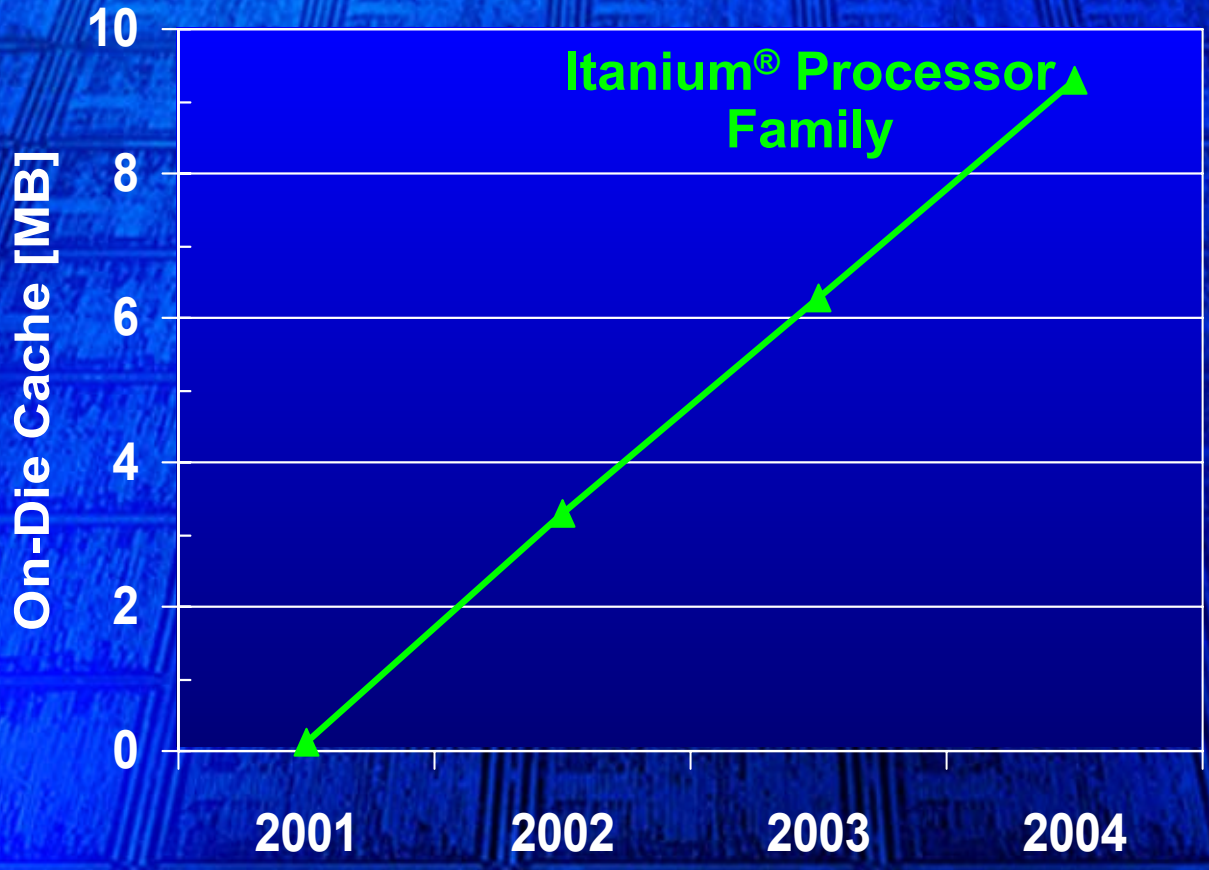
*From Itanium®2 processor to Montecito: 2-4X Performance**

intel

Source: Intel data, IBM.com, Sun.com

On-Die Cache Trends

- Significantly better latencies for on-die cache
- Higher cache bandwidth on-die (48GB/s for Madison)



intel

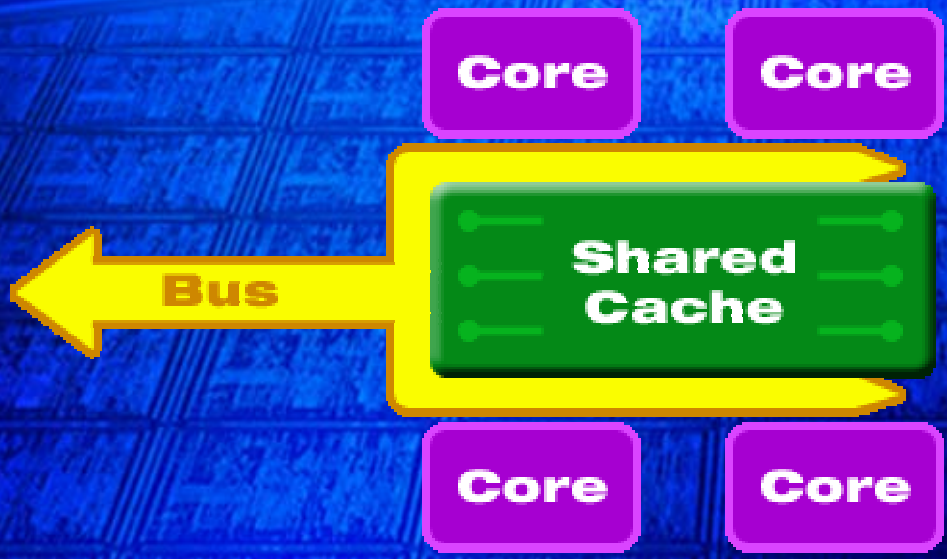
On-die Cache = Superior Latency & Bandwidth

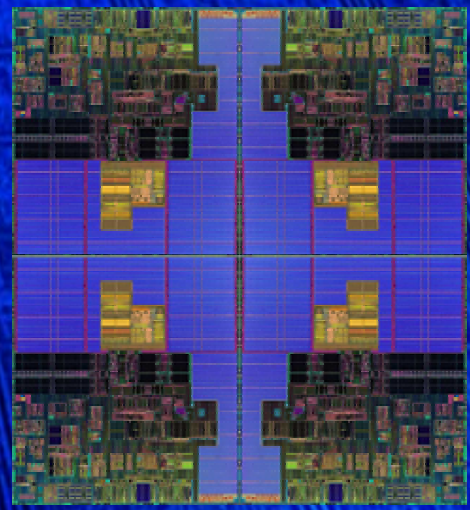
Future Itanium Processor Family Technologies

- Higher frequencies
- Larger caches
- Lower power implementations
- Multi-core technology
- Multi-threading
- Faster system bus
- Additional RAS technologies

Multiple new technologies to exploit Increasing transistor counts

A Simple Extrapolation





4 Processor system on a chip, Integrating:

- 4 Itanium® 2 processor Cores ~120 M transistors
- Shared Cache 16 MB ~900 M transistors
- Leaf interconnect

1B Transistors Possible in 65 nm process with < 500 sq mm die size

intel

Itanium® 2 Processor Baseline

Itanium® 2 Processor Core ~30 M transistors

L3 Cache ~65 M transistors per MB

	1 Core	2 Cores	4 Cores
3MB Cache	~ 225 M	~ 255 M	~ 315 M
6 MB Cache	~ 420 M	~ 450 M	~ 510 M
9 MB Cache	~ 615 M	~ 645 M	~ 705 M
12 MB Cache	~ 810 M	~ 840 M	~ 900 M
18 MB Cache	~ 1.2 B	~ 1.23 M	~1.3 B

Several paths to 1B Transistors

CMP Challenges

- How much Thread Level Parallelism is there in non-embarassingly parallel workloads?
- Ability to generate code with lots of threads & performance scaling
- Thread synchronization
- Operating systems for parallel machines
- Single thread performance
- Power limitations
- On-chip interconnect infrastructure
- Off-chip interconnect infrastructure
- Memory and I/O bandwidth required

Design Challenges for 1 B Transistors

- Design Complexity
  - Productivity Tools and Methods Advance
    - …But at slower rate than Moore's Law
    - Replicating cores improves productivity
- Visibility for Test & Debug
  - Pin Bandwidth/Transistor continues to decline
  - Shrinking dimensions, increasing speeds, …
- Power
  - Power Delivery – di/dt of Amps/nano-second
  - Thermals: Overall power and thermal density

Outline

- Semiconductor Technology Evolution
- Moore's Law Video
- Parallelism in Microprocessors Today
- Multiprocessor Systems
- The Path to Billion Transistors
- Summary

Summary

- One billion transistors feasible within the next 4 years
- Chip Level Multiprocessing and large caches will get us there
- Semiconductor technology will continue to drive higher performance
- Itanium® 2 processor achieving record results today
- Madison extends performance by 30%-50% in 2003
- Innovation for Itanium processor family occurring on multiple vectors
  - Multi-core Montecito planned for 2005
  - Multi-threading
  - Higher frequencies
  - larger caches
  - Faster buses